© 2020. Published by The Company of Biologists Ltd | Development (2020) 147, dev184481. doi:10.1242/dev.184481

Endogenous CRISPR/Cas9 arrays for scalable whole-organism lineage tracing

James Cotterell[1,2,*], Marta Vila-Cejudo[2], Laura Batlle-Morera[2] and James Sharpe[1,2,3,*]

ABSTRACT

The past decade has seen a renewed appreciation of the central importance of cellular lineages to many questions in biology (especially organogenesis, stem cells and tumor biology). This has been driven in part by a renaissance in genetic clonal-labeling techniques. Recent approaches are based on accelerated mutation of DNA sequences, which can then be sequenced from individual cells to re-create a 'phylogenetic' tree of cell lineage. However, current approaches depend on making transgenic alterations to the genome in question, which limit their application. Here, we introduce a new method that completely avoids the need for prior genetic engineering, by identifying endogenous CRISPR/Cas9 target arrays suitable for lineage analysis. In both mouse and zebrafish, we identify the highest quality compact arrays as judged by equal base composition, 5′ G sequence, minimal likelihood of residing in the functional genome, minimal off targets and ease of amplification. We validate multiple high-quality endogenous CRISPR/Cas9 arrays, demonstrating their utility for lineage tracing. Our pragmatically scalable technique thus can produce deep and broad lineages *in vivo*, while removing the dependence on genetic engineering.

KEY WORDS: CRISPR, Human, Lineage Tracing, Mouse, Zebrafish

INTRODUCTION

Development describes the process whereby a single totipotent zygotic cell transforms into a complex multicellular organism. Defining the early patterns of cell division in developing organisms is of paramount importance to understand and ultimately control the mechanisms of cell fate decisions that impact on developmental, stem cell and cancer biology. The traditional method for defining the early patterns of cell division focused on fate mapping, which, when performed at cellular resolution, is called lineage tracing (Vogt, 1929; Sulston et al., 1983; Clarke and Tickle, 1999; Stern and Fraser, 2001).

Original methods for labeling cells depended on direct injection of a chosen cell early in development, with dyes or enzymes that would be retained in daughter cells over multiple rounds of division (Weisblat et al., 1978). A major improvement was the introduction of genetic methods, which removed the need for physical manipulation of the embryo. These relied on a stochastic molecular event permanently activating expression of a marker, which would be clonally inherited by all daughter cells of the cell of origin [e.g. the LacZ transgene (Soriano, 1999) or GFP transgene (Mao et al., 2001)]. These 'single-label' methods, however, could not analyze multiple clones in the same piece of tissue, and were subsequently superseded by the various 'rainbow-label' techniques in which the engineered stochastic genetic events activated random combinations of different fluorescent proteins (Livet et al., 2007), thus allowing the labeling of many different clones with multiple different colors (Fig. 1A).

Clonal analysis distinguishes those cells within the clone from those outside, but involves only a single genetic event per clone, and thus cannot contain rich information about the longer history of the cells. Any approach to recreate genuine cellular lineage trees from the end-point cells requires the recording of multiple successive genetic changes within the same cell over time (Fig. 1B). The earliest attempts to use somatic mutation to generate cellular lineage trees focused on microsatellite mutations that act as 'molecular tumor clocks' that recorded past tumor histories (Shibata et al., 1996). This type of approach using the genomic variability within an organism to elucidate the cell lineage tree has been described as 'phylogenetic fate mapping' (Frumkin et al., 2005; Salipante and Horwitz, 2006).

This approach has been developed further to define cellular lineage relationships using genetically engineered mice whose DNA-repair systems were compromised, resulting in more mutations at the 120 MS loci analyzed (Frumkin et al., 2008). Such mice have a high rate of mutations in MS loci, and develop a variety of spontaneous tumors (Prolla et al., 1998). This accelerated the accumulation of mutations, thus reducing the amount of sequencing required and allowing the first lineage trees to be derived.

The recent advent of CRISPR technology (Doudna and Charpentier, 2014) has provided an alternative method for producing multiple independent mutations within the same cells. The targeted nature of CRISPR allows mutations to be targeted to a compact region of the genome called an array [instead of the 120 microsatellites used in by Frumkin et al. (2008), for example] that can be readily deep sequenced. These approaches offer scalability to whole-organism lineage tracing as each CRISPR/Cas9 target can potentially encode at least a single bit of information. Therefore, the total amount of information encoded by an array of CRISPR/Cas9 sites is 2^n where n is the number of CRISPR/Cas9 target sites in the array. For lineage tracing, the amount of information encoded by the array should be higher than the number of cells in the tissue that we want to lineage trace. To perform lineage tracing in a whole-mouse embryo of 12 billion cells, 33 CRISPR target sites in an array (or multiple arrays) is the theoretical minimum that would be required to provide enough diversity.

Multiple approaches have employed CRISPR for lineage tracing *in vivo*. One such approach, named GESTALT, focuses on the

[1]European Molecular Biology Laboratory (EMBL) Barcelona, 08003 Barcelona, Spain. [2]Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), 08003 Barcelona, Spain. [3]Institucio' Catalana de Recerca i Estudis Avancats (ICREA), 08010 Barcelona, Spain.

*Authors for correspondence (james.sharpe@embl.es; james.cotterell@embl.es)

J.C., 0000-0003-4251-6382; M.V., 0000-0002-3138-2668; J.S., 0000-0002-1434-9743

Handling Editor: James Briscoe
Received 2 September 2019; Accepted 6 March 2020

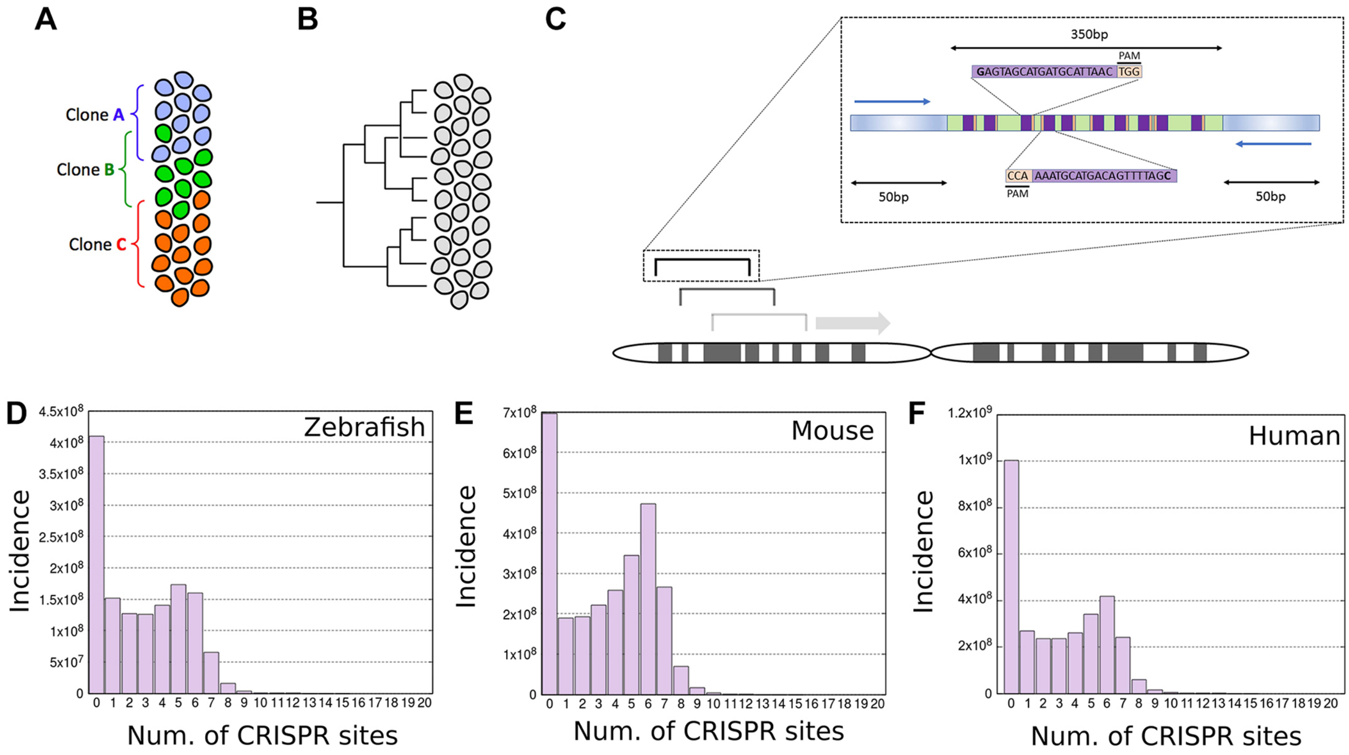

Fig. 1. Identification of endogenous CRISPR arrays. (A) Lineage dendrogram that can be generated with traditional genetic recombination approaches to lineage tracing. (B) Lineage dendrogram that can be generated with whole-organism lineage-tracing approaches. (C) The zebrafish, mouse and human genomes were scanned using a moving window of 450 bp (inset). The flanking 50 bp regions (shaded blue) are reserved for PCR amplification primer identification. We identify clusters of non-overlapping CRISPR/Cas9 targets within those windows (i.e. a 20 bp distance was required between adjacent NGG PAM sites and adjacent CCN PAM sites, a 0 bp distance was required between a NGG PAM site followed by a CCN PAM site, and a 40 bp distance was required between a CCN PAM site followed by a NGG PAM site). No additional filters from the two pipeline versions were used. CRISPR/Cas9 target sites are illustrated by the purple box and the PAM sequences are shown in the orange box. Primers are illustrated by the blue arrows. (D-F) Histograms of the incidence (*y*-axis) of 350 bp windows containing specific numbers of CRISPR/Cas9 target sites (*x*-axis) in the zebrafish (D), mouse (E) and human (F) genomes.

generation of a synthetic compact array of CRISPR/Cas9 targets that was introduced into the genome of zebrafish (McKenna et al., 2016). Zygotic injection of the CRISPR machinery that targets that array therefore generates diversity at that location, which can be readily deep sequenced and used for lineage tracing. A second approach generates lineaging barcodes by targeting the same sequence in single or multiple repeats of a transgenic fluorescent protein gene (Junker et al., 2017preprint; Schmidt et al., 2017; Alemany et al., 2018; Spanjaard et al., 2018). However, both of these approaches suffer from the drawback of requiring the generation of a transgenic animal. Recently developed Tracerseq (Wagner et al., 2018) can be used on wild-type embryos but can only be used in conjunction with single cell sequencing platforms as it requires barcodes to be sequenced that have integrated into different regions of the genome.

Here, we set-out to discover whether a practical method of CRISPR-based lineage analysis could be achieved without having to genetically engineer the genome in advance: in particular, whether endogenous sites within the genome could act as suitable CRISPR/Cas9 arrays for this task. This approach would have the advantage of functioning on wild-type embryos simply by injecting the CRISPR machinery into the one-cell zygote (or later stage). CRISPR/Cas9 target sites are constrained by the requirement for a proto-spacer adjacent motif (PAM) that has the form NGG for Cas9 (or CCN on the opposite DNA strand) (Mojica et al., 2009; Shah et al., 2013; Jinek, 2012; Sternberg et al., 2014). Because a GG or CC dinucleotide is expected to arise on average every eight base pairs, we reasoned that, by chance, arrays of compact CRISPR/Cas9 targets should appear naturally in most large genomes (i.e. >1 Gb). We explain the criteria employed to find suitable CRISPR/Cas9 arrays, illustrate our findings for zebrafish and mouse genomes, and validate *in vitro* two of those arrays demonstrating that the target sites are efficiently edited as expected and that the method can indeed be used for lineage tracing.

RESULTS

To search for suitable endogenous CRISPR/Cas9 arrays, we obtained appropriate genomic regions for zebrafish, mouse and human from the UCSC genome browser (see Materials and Methods). We focused on regions of the genome that could be constructed with paired end reads on the Illumina Miseq, as this platform offers an appropriate balance between paired-end read length (up to 2×250 bp for the version 2 kit) with maximal throughput (number of reads). A small amount of overlap between the paired-end reads allows for the region to be reconstructed and here we allow 50 bp of overlap. Hence, we searched the genome using a conservative 450 bp moving window (Fig. 1C) that would allow efficient sequencing of clusters using a Miseq version 2 or 3 reagent kit. We reserve the first and last 50 bp of the sequence for primer identification in order to amplify the region via PCR. This resulted in a window of 350 bp for searching for the maximal number of CRISPR/Cas9 target sites.

We searched for PAM sequences (NGG on the sense strand or CCN on the antisense strand) throughout the genomes of zebrafish, mouse and human. Overlapping target sites suffer from the drawback that editing events at one of the overlapping sites are likely to destroy the target site of another overlapping site, thus reducing the potential variability that can be generated in the CRISPR/Cas9 array. Therefore, in order to maximize the variation amongst our CRISPR/Cas9 targets, we focused on non-overlapping CRISPR/Cas9 target sites (Fig. 1C inset) by searching for PAM sequences that had a space of at least 23 bp between them (only 3 bp between a GG followed by a CC as the target sequence would be read in opposite directions from the PAM). Histograms of the frequency of windows with different numbers of non-overlapping CRISPR/Cas9 targets sites with 5′G nucleotides are shown in Fig. 1D-F for the zebrafish, mouse and human genomes, respectively. As expected, there are a huge number of windows with many CRISPR/Cas9 target sites across all three genomes, as predicted from the frequency of CC and GG dinucleotides. There is a peak in window frequency for windows containing five non-overlapping target sites for the zebrafish genome and six non-overlapping target sites for both the mouse and human genomes (Fig. 1D-F, respectively).

This high frequency of windows with many non-overlapping CRISPR/Cas9 target sites allowed us to use very stringent selection criteria so that we could focus on identifying the best possible CRISPR/Cas9 arrays. We therefore set up a series of selection criteria with the aim of applying them in order from the most stringent to the least stringent to minimize computation. We defined a CRISPR/Cas9 array as a contiguous region of the genome with more than eight CRISPR/Cas9 sites per 350 bp window. The first version of our pipeline is described in the supplementary Materials and Methods (section S1 and Tables S1-S3). This was superseded by version 2 of our pipeline where we applied the following filters: (1) balanced base frequency filter; (2) minimal off-target filter; (3) *in silico* CRISPR/Cas9 activity prediction filter; (4) non-functional site filter; and (5) filter for arrays containing SNPs. It also includes an optimizing function to select for the configuration giving the greatest number of non-overlapping target sites in the array.

We chose to focus on windows with a balanced base frequency, as we deemed that the higher information content would remove the likelihood of secondary structure in the sgRNA (see Discussion). Therefore, we set a balanced base frequency filter such that the base with the highest frequency (count in window) could not be more than 50% higher than the base with the lowest frequency. We employed a filter that selects for endogenous CRISPR/Cas9 arrays with minimal off-targets because off-targets potentially cause detrimental effects on the organism and also potentially quench the activity of Cas9, therefore leading to less efficient editing of our region of interest. Therefore, for each potential sgRNA target array we created a Bowtie2 (Langmead and Salzberg, 2012) query file consisting of the set of target sequences. We then used Bowtie2 to search a prebuilt index of the corresponding genome. We set the Bowtie2 options so that only the two highest scoring alignments are reported for each sgRNA with sequences allowed to differ by up to 1 bp. Arrays only passed the filter if no similar sequence could be found in the genome for any of the sgRNAs in the array. The non-functional site filter also employed in both pipeline versions was used to restrict endogenous arrays to those that were less likely to have detrimental functional effects of the organism. We removed those that reside in a coding sequence or upstream regulatory region (up to 5 kb upstream). These are defined in the UCSC upstream5000 and mRNA files for mm10 and danRer10. We combined these putative functional sequences and built an index using bowtie2 for each species. We then used bowtie2 to search for the entire window of interest (450 bp) in the functional sequences for that respective species. If the sequence was not found among the functional sequences then the array passed the filter. The *in silico* CRISPR/Cas9 activity prediction filter uses sgRNAScorer2.0 (Chari et al., 2017). The aim was to select arrays where all CRISPR/Cas9 target sites were predicted to have high and similar editing efficiency. The filter therefore selected only those arrays where all CRISPR/Cas9 targets in the array were in the top quartile for their predicted activity score. We also employed a filter to find arrays that were likely to harbor a single nucleotide polymorphism (SNP) in order to use these SNPs to work out which allele a read derived from by using data from dbSNP (see Materials and Methods). Finally our algorithm attempts to find the optimal configuration of non-overlapping target sites within any given window by employing a triangular distance matrix such that the configuration with the most number of non-overlapping CRISPR/Cas9 target sites per window is selected (explained in more detail in supplementary Materials and Methods, section S2 and Fig. S1). All identified arrays were suggested to be amplifiable by PCR using the 50 bp flanking sequences and primer3 (Untergasser et al., 2012).

The final distribution of endogenous CRISPR/Cas9 arrays over the genomes when searching for arrays with more than eight CRISPR/Cas9 target sites per window is shown in Fig. 2A-C for zebrafish, mouse and human, respectively. As can be seen, potential CRISPR/Cas9 arrays exist on almost all chromosomes, offering a flexible choice of targets. We provide a full list of the coordinates of endogenous CRISPR/Cas9 arrays (10 sites per array) for each species in bed format in the supplementary Materials and Methods (section S3). The results of the number of endogenous CRISPR/Cas9 arrays passing all filters for all three species when we search for more than CRISPR/Cas9 target sites are described with the inset nested Venn diagrams. The number of endogenous CRISPR/Cas9 arrays passing all filters can be seen to rapidly diminish when increasing the number of CRISPR/Cas9 target sites that are required per array.

We validated one target array in each of the zebrafish (Z4 array identified with version 1 pipeline), mouse (M7 array) and human (H5 array) genomes (the oligonucleotide sequences to produce the corresponding sgRNAs are given in Tables S7-S10). For targeting in zebrafish, we transcribed the 10 sgRNAs *in vitro* and individually microinjected them into the one-cell stage zebrafish yolk sac with CAS9 protein with a nuclear localization signal (Fig. 3A). Genomic DNA was extracted from embryos 30 h later. We cloned the targeting sequences for the mouse and human arrays into PX458 (Ran et al., 2013), which expresses both the sgRNA and Cas9 (Fig. 3B). We transfected the resulting 10 mouse M7 targeting vectors individually into mouse 3T3 cells and the 10 human H5 targeting vectors individually into human HEK293 cells. 65 h later, the cells were sorted for GFP expression and genomic DNA was extracted from the positive fraction. We then amplified the Z4, M7 and H5 arrays (shown with their spacer regions and PAM sequences in Fig. 3C-E, respectively) and performed the surveyor nuclease assay.

The surveyor nuclease assay can detect the presence of mutations in a DNA fragment of interest if a wild-type reference exists. In the surveyor nuclease assay, potentially mutated DNA is mixed with reference wild-type DNA. The two are melted and annealed, resulting in hybrid double-stranded DNA with bulges if a mutation was present. Surveyor nuclease is then added, which cuts hybrid DNA at these bulges of mismatched DNA. The resulting DNA is then run on a gel and two fragments are detected, the sizes of which correspond to the position of the mismatch relative to the ends of the DNA fragment if a mutation was present. CAS9 typically introduces indels in DNA 3 or 4 bp upstream of the PAM site, allowing us to

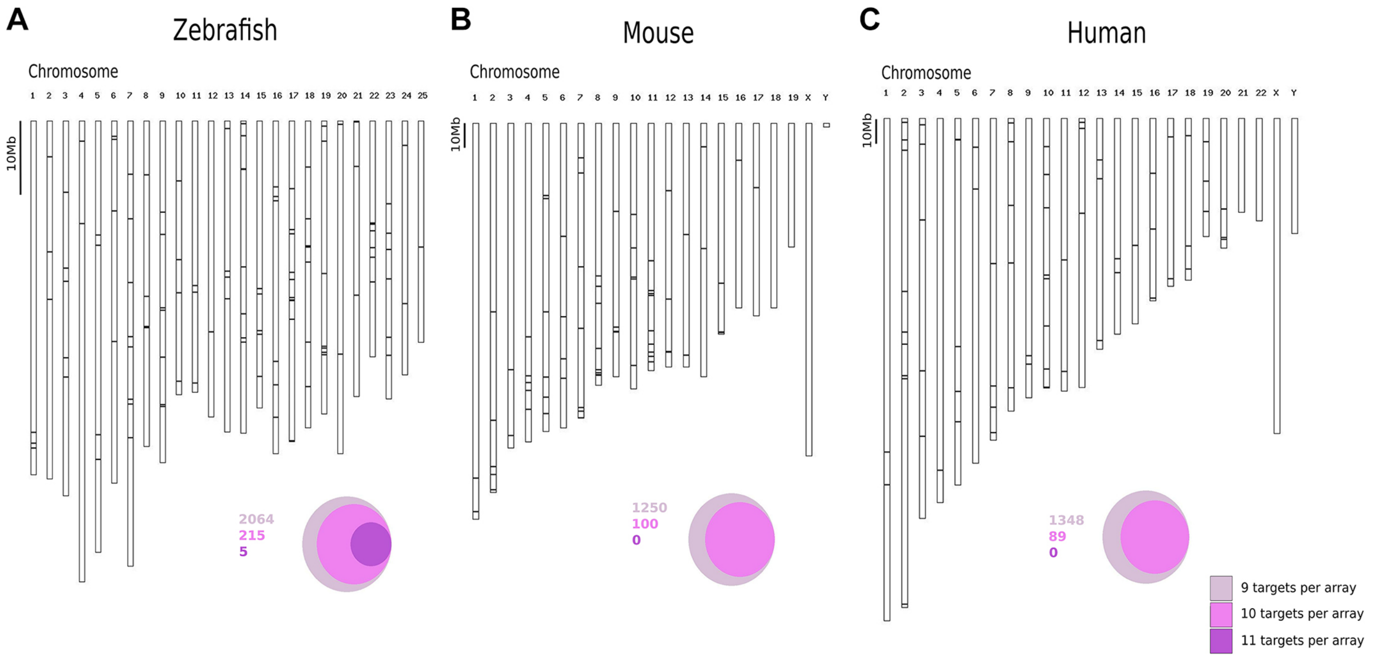

Fig. 2. Summary of CRISPR cluster regions of the zebrafish and mouse genomes. (A-C) Distribution of the filtered endogenous CRISPR/Cas9 target arrays containing more than eight targets over the zebrafish, mouse and human genomes using version 2 of the array identification pipeline. Individual CRISPR/Cas9 arrays are represented by a black line. The number of CRISPR/Cas9 arrays with different numbers of targets are shown by the inset Venn diagrams for each species. Areas represent the log of the number of arrays with a given number of target sites in the nested Venn diagrams and the numbers are given in the respective colors.

predict the fragment sizes produced in the surveyor nuclease assay for each of our target sites. The expected fragment sizes depend on the location of the PAM site in the array. More peripheral PAM sites produce a small and a large fragment, whereas the most central PAM sites result in two similar-sized fragments. The resulting spectrum of fragments resembles an X shape for all arrays tested (Fig. 3F-H). The fragment sizes were then measured using a high-sensitivity DNA chip and the results are shown in Fig. 3I-K. This result confirms that all of our targets produce DNA fragments of the expected size (although three of the Z4 bands are faint). Therefore the endogenous CRISPR/Cas9 arrays that we have identified are genuine functional arrays that can be used for lineage tracing.

To interrogate the editing ability of our targeting vectors/sgRNAs in a quantitative and more-detailed fashion, we performed deep sequencing of the target arrays using the Miseq (Illumina). Furthermore, we aimed to detect any interference between the targeting sgRNAs by assaying them as pools rather than individually. Therefore, we transfected either the pool of the 10 M7 mouse targeting vectors or control empty PX458 vector into mouse 3T3 cells in triplicate. We also transfected either the pool of the H5 human targeting vectors or control empty PX458 vector into human HEK293 cells in duplicate. 65 h later, the cells were sorted for GFP expression and genomic DNA was extracted from the positive fraction. For targeting in zebrafish, the 10 Z4 sgRNAs were microinjected as a pool into the one-cell stage zebrafish yolk sac with CAS9 protein with a nuclear localization signal with a corresponding negative control (Fig. 3A). Genomic DNA was extracted from embryos 30 or 48 h later. We then PCR amplified the respective arrays (shown with their spacer regions and PAM sequences in Fig. 3C-E, respectively) and added on indexed (each sample was given a different index) Illumina adaptors with a further round of PCR. We performed deep sequencing using 2×250 bp or 2×300 bp cycles on the Miseq (see Materials and Methods). We applied a number of bioinformatic filters to removed sequencing errors and erroneous reads potentially generated by PCR flip-over [see Materials and Methods, supplementary Materials and Methods (section S4) and Fig. S2]. We identified the midpoint of each of the indels and calculated a histogram of editing incidence over the corresponding regions. The results shown in Fig. 3L (Z4), 3M (M7) and 3N (H5) confirm that peaks of more-frequent editing occur where we expect them (PAM position minus 3 bp as represented by the dashed vertical lines) and that very few edits are detected in the corresponding negative control samples (Fig. 3O-Q). Furthermore, there does not appear to be significant interference between targeting sgRNAs as all peaks are detected where they are expected. Finally, we analyzed the types of indels and the number of between-target dropouts that we generate for each of the sgRNA pools. The results are shown in Fig. S3 and show that significant target site drop out is only observed between the 4th and 5th targets in the H5 array using the H5 sgRNA pool. Taken together, these results show that indeed all sgRNAs are capable of editing their respective target site *in vitro* and *in vivo*.

In order to show that our tool could indeed be used for lineage tracing, we microinjected one-cell stage zebrafish embryos with CAS9 and an equimolar pool of the Z4 targeting sgRNAs and extracted genomic DNA 48 h later (Fig. 4A). We then PCR amplified the respective arrays and performed deep sequencing with the Miseq as described above. We analyzed the zebrafish embryo data with three different aligners, bowtie2 (Langmead and Salzberg, 2012), Bwa mem (Li, 2013 preprint) and Needleall (Rice et al., 2000), and confirmed that they identify indels in the same locations with different parameter sets [specifically, the gap open and gap extension penalties can have a major impact of indel calling; see Materials and Methods, supplementary Materials and Methods (section S5) and Fig. S4]. To generate the high confidence dataset for the zebrafish embryos, we only included reads where the

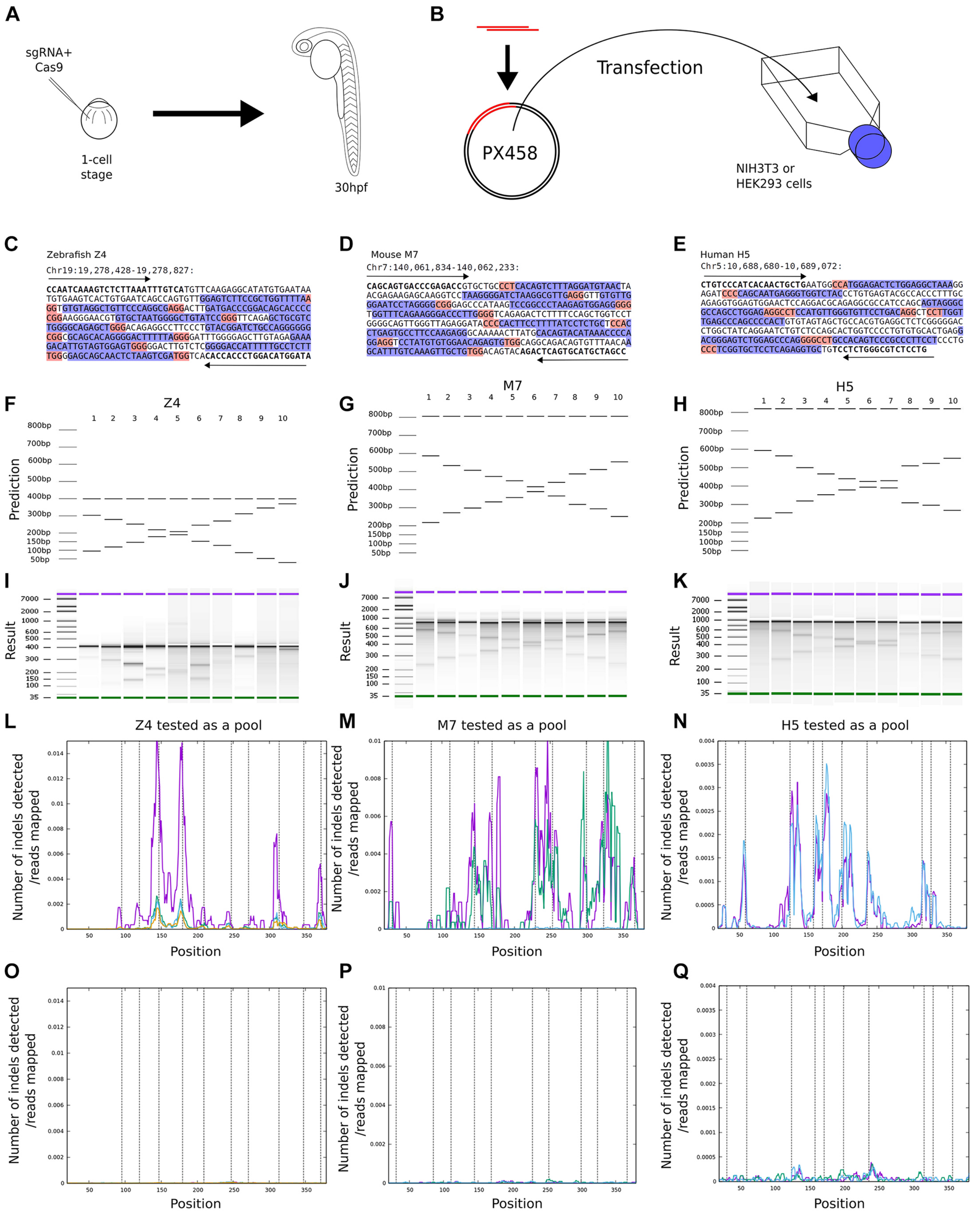

Fig. 3. See next page for legend.

Fig. 3. Validation of one CRISPR/Cas9 array from each of zebrafish, mouse and human. (A,B) Protocols for validation of CRISPR/Cas9 targets *in vivo* in zebrafish embryos (A) or *in vitro* in mouse or human cell lines (B). (C-E) Examples of endogenous CRISPR/Cas9 arrays from zebrafish (C), mouse (D) and human (E). The Primer3 PCR primer sequences of the arrays are shown in bold. The CRISPR/Cas9 targets sites are shown in blue. The PAM sequences are highlighted in red. (F-H) Fragment size predictions from a surveyor nuclease assay are shown for the three tested endogenous CRISPR/Cas9 arrays. The band at the higher molecular weight is the uncut amplicon and the two bands at lower molecular weight are the cut fragments. (I-K) Gels from a high-sensitivity DNA chip after application of the surveyor nuclease assay. Bands are of the expected size for all targets with sufficient signal. (L-N) Indel detection using Miseq deep sequencing of the zebrafish, mouse and human amplicons (showing three replicates for zebrafish and mouse samples, two replicates for human). The different colored lines show the number of indels (center point of indel and averaged over a window length of 5 bp) detected at that specific position in the amplicon for each of the different experimental repeats. Vertical dashed lines represent the expected positions of indels (3 bp downstream of PAM site). (O-Q) The negative controls for the respective samples in L-N (showing three replicates for each sample). The bowtie2 aligner was used for these samples with default parameters (i.e. rdg=5,3 and rfg=5,3).

spectrum of indels reported was consistent between the all three aligners [see Materials and Methods and supplementary Materials and Methods (section S5)]. The indels reported are generally small (<20 bp) in these data with only a small proportion of deletions spanning adjacent edit sites (see Figs S5-S7, each with a different alignment parameter set). The spectrum of indels in each read was then used to construct a dendrogram using the commonly used phylogenetic software PAUP (Swofford, 2017) [the full method is described in the Materials and Methods, and supplementary Materials and Methods (section S6)]. As the goal of our work here was to demonstrate a novel a lineaging tool and not to define any specific lineage, we used PAUP in neighbor-joining mode. Neighbor joining is not as accurate as other methods, such as maximum parsimony or maximum-likelihood, but it is less computationally intensive and comparison between datasets is more meaningful owing to its simplicity.

In order to generate a representative metric of the utility of the approach, we explored the distribution of the number of mutations of all reads. The histograms of dendrogram depth for the Z4 array are shown in Fig. 4B and Figs S9-S11, demonstrating that the full trees are generated by reads with up to eight mutations (mean number of mutations per allele is 3.03 for the zebrafish embryo in Fig. 4). The full dendrogram for one zebrafish embryo consisting of 1572 alleles is shown in Fig. 4C and a subset of that dendrogram is shown in Fig. 4D. To explore the consistency of this approach, we deep sequenced a total of 10 embryos microinjected with the Z4

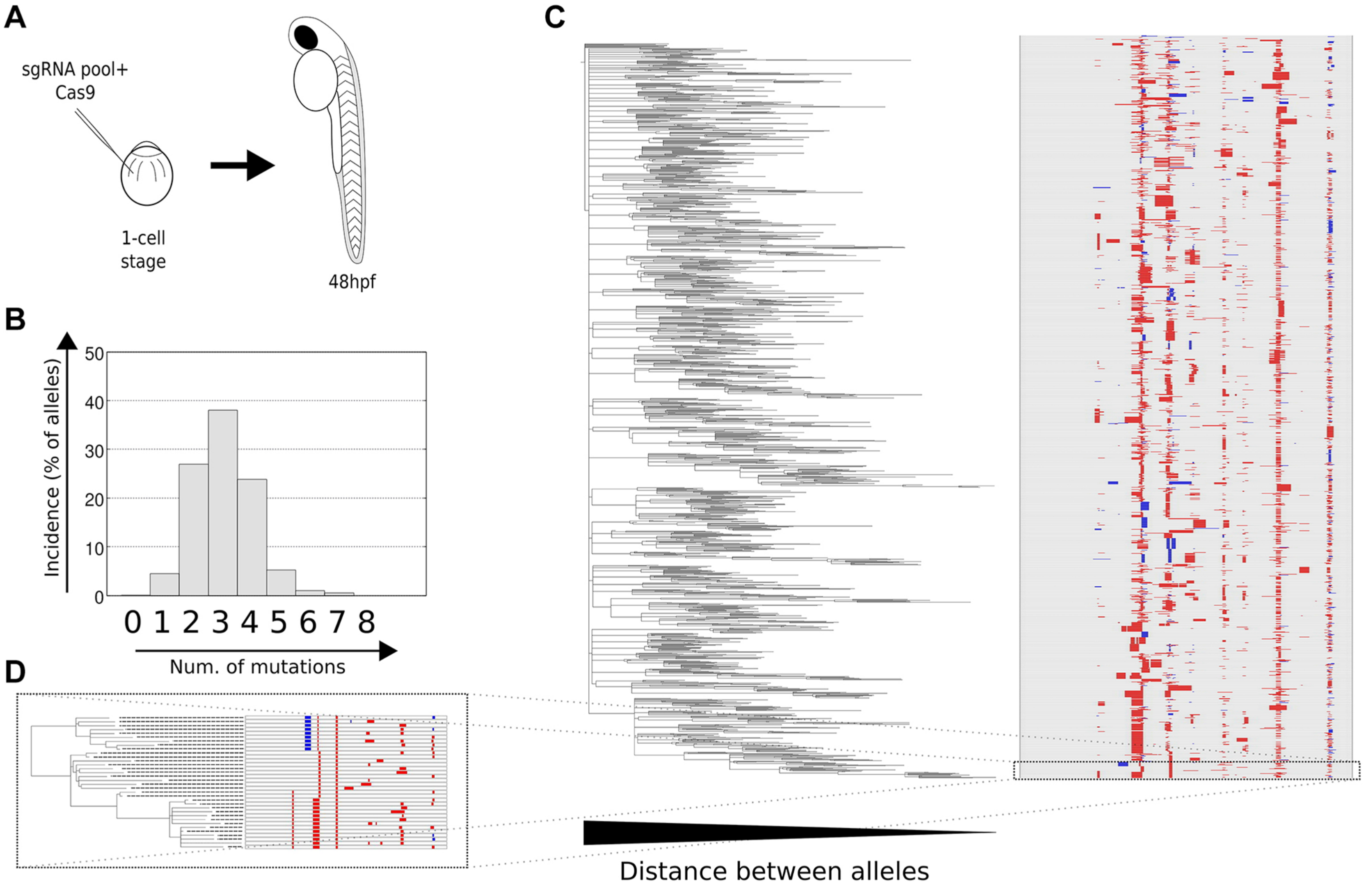

Fig. 4. The approach can be used to perform lineage tracing in zebrafish embryos. (A) Schematic showing the approach to performing lineage tracing in the zebrafish. (B) Histogram of the number of features per allele in the zebrafish 48 hpf embryo 4 dataset. (C) Full lineage tree for the zebrafish embryo index 4 built using indel spectrum information of amplicons after deep sequencing. This tree consists of 1572 leaf and zero-length edges are collapsed. (D) A sub-branch of the full lineage. Red blocks indicate a deletion and blue blocks indicate an insertion with the size of the block representing the size of the indel. Only reads where all three aligners reported the same CIGAR string (indel spectrum) were used to construct the zebrafish lineage trees. For the histogram in B and the lineage tree in C, the aligners used a gap open penalty of 5 and gap extension penalty of 3 – i.e. Alignment parameter set 1 – see Materials and Methods. See Materials and Methods and Figs S9-S11 for alternative trees built from other zebrafish embryos and aligner parameter sets.

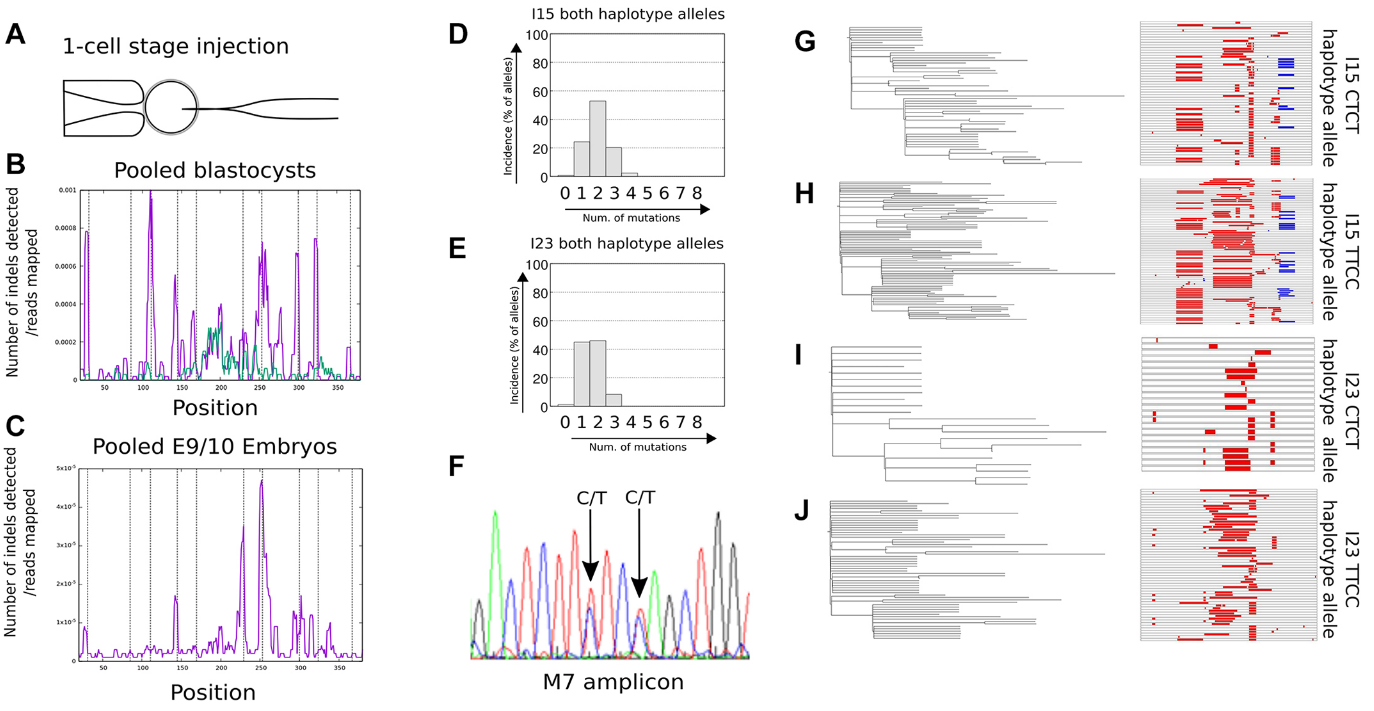

Fig. 5. The approach can be used to perform lineage tracing in mouse embryos. (A) Schematic showing the approach to perform lineage tracing in the mouse. (B) Indel detection using Miseq deep sequencing of the M7 amplicon from pooled blastocysts (purple line) or negative control (turquoise line, uninjected pooled blastocysts). (C) Indel detection using Miseq deep sequencing of the M7 amplicon from pooled E9.0-E10 embryos. For both B and C, the number of indels (center point of indel and averaged over a window length of 5 bp) detected at that specific position in the amplicon is shown. Vertical dashed lines represent the expected positions of indels (3 bp downstream of PAM site). The bowtie2 aligner was used for the indel position data with default parameters (i.e. rdg=5,3 and rfg=5,3). (D,E) Histograms of the number of features per allele in the mouse embryos datasets with Illimuna indices 15 (D) and 23 (E). (F) A Sanger sequencing trace showing the two single nucleotide polymorphisms in the M7 array of CBA/Bl6 mice used to split mapped.sam data into two separate datasets. (G,H) The lineaging result from an individual mouse embryo (Illumina index 15). The full dendrogram can be split into two, using the two SNPs that are present in this array in the CBA/Bl6 background. (I,J) As in G,H, but for a mouse embryo with an Illumina index of 23. Data for other heterozygous embryos can be found in Fig. S12. The bwa mem aligner was used to generate the datasets for these lineage trees and mutation histograms with parameters O=6, E=1 and w=400 (all other parameters are as default).

sgRNA pool with genomic DNA extracted at either 30 or 48 h post fertilization. The resulting dendrograms are shown in Figs S9-S11 (each with a different alignment parameter set). This result demonstrates that we can consistently generate edits across the endogenous Z4 CRISPR/Cas9 array. Taken together, these data demonstrate how our approach can be used effectively to perform lineage tracing in embryos.

Finally, we attempted to use our M7 array to perform lineage tracing in mouse embryos. We microinjected one-cell stage embryos with CAS9 with a nuclear localization signal and an equimolar pool of the M7 targeting sgRNAs (Fig. 5A). Embryos are injected at the one-cell stage with the pool of M7 sgRNAs and Cas9, and either developed until 4.5 days (blastocysts) or transferred into pseudopregnant CD1 mothers and allowed to develop until the E9.0-E10 stage. The number and percentage of embryos surviving to the blastocyst stage and the number and percentage of embryos recovered (embryos judged to have a gross normal morphology) at E9-E10 after embryo transfer is summarized in Table 1. These numbers are in line with those of other groups (Wang et al., 2013), indicating that CRISPR/Cas9 components do not perturb normal mouse development.

Genomic DNA was extracted from pooled blastocysts (46 injected with M7 pool and five uninjected control) or E9.0-E10 stage embryos, the M7 array PCR amplified and deep sequencing performed with the Miseq and bioinformatics in a similar manner to that described for zebrafish (see Materials and Methods). We quantified the number of indels over the array for the pooled blastocyst and pooled E9.0-E10 stage embryos. For pooled blastocysts, peaks are found at all but one of the expected locations (dashed lines in Fig. 5B). For the pooled embryos, however, three or four of these peaks appear to have been lost (Fig. 5C). The histograms of number of mutation per allele are presented for two embryos in Fig. 5D,E (Illumina index 15 and 23), demonstrating that this technique generates significantly fewer edits per allele in the mouse than in the zebrafish.

The M7 array contains two SNPs, T/C and C/T, two base pairs apart, as shown by the Sanger trace in Fig. 5F. We split the resulting mapped reads into separate datasets depending on whether they were derived from the TTCC haplotype allele or the CTCT haplotype allele. We defined seven out of the 18 embryos that we deep sequenced as heterozygote based on the highest frequency allele contributing <75% of the reads [see Materials and Methods, supplementary Materials and Methods (section S7) and Table S6]. We then performed lineaging and constructed two dendrograms for

Table 1. Summary of survival and recovery of injected mouse embryos

	Number of embryos microinjected	Number of embryos surviving to blastocyst stage (%)	Number of blastocysts transferred	Number of embryos recovered at E9.0-E10 (%)
Batch 1	96	90 (94%)	28	12 (43%)
Batch 2	59	49 (83%)	49	17 (35%)

each haplotype allele in the same way as described previously for the zebrafish (shown in Fig. 5G-J for the two embryos with most unique alleles and Fig. S12 for all seven heterozygous embryos). This result serves as a proof-of-principle that we can use these endogenous arrays to perform lineage tracing in mouse and that we can determine in which allele an edit resides using single-nucleotide polymorphisms.

DISCUSSION

By identifying endogenous CRISPR/Cas9 arrays with appropriate properties, we demonstrate that it is not necessary to generate a transgenic animal to use CRISPR/Cas9 arrays for lineage tracing, as in previous studies (McKenna et al., 2016; Junker et al., 2017 preprint; Schmidt et al., 2017; Alemany et al., 2018; Spanjaard et al., 2018). The approach described here does not depend on single-cell sequencing and can be easily extrapolated to any other species of interest. The NGG PAM is essential for CRISPR/Cas9 target sites. Given that a CC or GG dinucleotide is expected to occur on average every 8 bp, it is not surprising that we find a huge number of CRISPR/Cas9 target arrays throughout the zebrafish and mouse genomes. This large number of potential sites allows us to use strict selection criteria for optimal target site features, resulting in the highest quality potential sites.

Here, we have demonstrated that our technique can generate near-saturating edits with corresponding deep dendrograms in zebrafish, and we have basic proof-of-principle in mouse. The indel data used to generate the mouse dendrograms are less reliable than those from zebrafish as we have only used a single aligner rather than multiple aligners (with reads only included if the indel spectra is identical for each aligner). Furthermore, in order to fully explore the consistency of the lineage reconstruction between individuals, labels (cell type or spatial information, for example) for the leaf nodes are required that can be added in future by adapting our approach to a single cell sequencing pipeline (see below). Nevertheless, when we compare the mouse and zebrafish datasets, we find that variability in the number of edited alleles and edits per allele is evident in both species, which likely arises from variability in microinjection parameters. We find that we generate significantly fewer edits per allele in the mouse system than in the zebrafish, as can be seen by comparing the histograms in Figs 4B and 5D,E. As a consequence, the lineaging dendrograms are clearly less complex for mouse than for zebrafish, and homoplasy can be observed because there are some edits shared between different alleles and different sub-branches of the same tree. Therefore, improvements in the mouse system are required in order to increase the amount of edits reported per cell and increase the amount of lineage information.

Adaptations to the current approach may involve using new technologies to expand the number of potential endogenous sites that can be used for whole-organism lineage tracing. For example, new nuclease enzymes that have different PAM sequence specificities, such as cpf1, CasX and Y or Cas9 (enzymes that have been engineered to alter their PAM specificities), may increase the number of sites on offer (Zetsche et al., 2015; Kleinstiver et al., 2015; Burstein et al., 2017). The continued improvement in defining the most efficient sequences for targeting and minimizing off-targets also promises to refine our approach further (Wang et al., 2014; Doench et al., 2014, 2016; Xu et al., 2015). Furthermore, longer read sequencing technologies, such as Pacbio and Oxford nanopore, will result in deeper and broader lineaging dendrograms through the inclusion of more CRISPR/Cas9 sites and thus the generation of more bits of information. Other lineaging approaches have attempted to increase the number of available bits through self-targeting evolving barcodes (Kalhor et al., 2017, 2018) but they have the drawback that it makes it more difficult to reverse engineer the history of the barcode and thus construct the lineage tree. Future versions of the system described here could also use inducible elements allowing lineaging to be addressed more effectively at later stages of development. For example, the heat shock-inducible Cas9 and the sgRNAs could be encoded on a tol2 element containing plasmid and injected with transposase in a similar manner to that demonstrated with the scGESTALT approach, but instead performed on wild-type embryos (Raj et al., 2018).

Some groups have also recently used single cell sequencing alone at different developmental time points to construct pseudo-cellular lineages (Wagner et al., 2018; Briggs et al., 2018; Farrell et al., 2018). The combination of a lineage-tracing strategy such as that described here with single cell and spatial sequencing approaches will impact lineage tracing in multiple ways. Indeed, recently the use of single cell transcriptomics in combination with CRISPR/Cas9-based lineage tracing (scGESTALT) has been demonstrated in zebrafish to help uncover restrictions at the level of cell types, brain regions and gene expression cascades during differentiation (Raj et al., 2018). Although single cell sequencing approaches currently have a much lower throughput than standard deep sequencing approaches, they promise to improve in the future. Whole-organism lineage tracing in combination with single cell sequencing has also recently been used to aid in elucidating the mapping from progenitor cell to adult cell (Alemany et al., 2018). The endogenous array approach described here is also amenable to parallel lineage tracing and single cell transcriptomics if one applies the single cell genome and transcriptome sequencing (G&Tseq) pipeline (Macaulay et al., 2015) to first separate genomic DNA from mRNA for sorted single cells. Finally, the combination of genetic lineaging approaches with techniques to define where cells are in physical space will become increasingly common. One such instance of performing CRISPR-based lineaging tracing while defining where cells are in space has recently been demonstrated *in vitro* (Frieda et al., 2017). We envision that all three of these techniques, genetic engineering-based lineaging tracing, spatial and single cell genomics/transcriptomics, will be used combination to provide a rich plethora of information with which to address multiple questions in developmental and cancer biology.

It is possible that multiple alleles from the same cell could be sequenced when performing lineage tracing using endogenous CRISPR/Cas9 arrays in species that are multiploid. This could lead to an erroneous lineage tree being generated. Here, we have demonstrated SNPs as one way to circumvent this problem by giving information on which allele a read derives from. Other potential ways around this problem include targeting a sex chromosome or using single cell genomic sequencing. However, not all species have a clearly identifiable heteromorphic sex chromosome and single cell genomics is currently not as high throughput as deep sequencing bulk tissue. Furthermore, by using a SNP to split the reads into two alleles, one can build two dendrograms for each individual organism. These serve as internal experimental repeats that can be used to improve and validate the accuracy of the resulting dendrogram if the approach is amended to include additional label information such as single cell transcriptomic or spatial information, allowing one to anchor the datasets. Alternatively, when used in combination with a single cell sequencing approach to determine when reads derive from the same cell, our method potentially allows one to generate double the amount of lineage information, as the two alleles function as independent arrays.

An important property of any CRISPR-based whole-organism lineaging system is that it must be capable of generating enough information (bits) to uniquely identify different cells. Here, we show that our system is indeed capable of generating such diversity. The dendrogram shown in Fig. 4C consists of 1572 alleles, demonstrating that we can generate significant diversity on par with GESTALT. Previously, we have mentioned that 33 CRISPR target sites would be the theoretical minimum required to provide enough diversity to perform lineage tracing in the whole mouse embryo of 12 billion cells. However, in reality, many more sites are required owing to factors such as homoplasy, target site drop-out and the fact that the editing is not synchronized with cell division. Indeed, simulations of CRISPR/Cas9 lineage tree recorders have demonstrated that >100 targets would be required to accurately reconstruct even a ∼50 K cells tree (Salvador-Martínez et al., 2019). The fact that the system described here is easily scalable when used in a single cell sequencing pipeline (multiple arrays and both alleles can potentially be sequenced for each cell) to generate more lineage information is thus important.

In summary, there are multiple advantages of our endogenous CRISPR/Cas9 array approach over previous approaches. First, it can be used to study cellular lineage in species for which it is difficult or impossible to generate transgenic animal lines. In particular, we envision that this technique could be effectively used with human pre-embryos and human iPS/ES cell-derived organoids (with our *in vitro* validated H5 array and others). For example, recently, single cell resolution gene expression time-courses have been elucidated for human pre-implantation embryos (Stirparo et al., 2018). Although these data allow one to generate a pseudolineage, defining which cell type populations are the genuine progenies of previous cell type populations requires a lineaging tool, such as the one we describe here. Second, even when it is possible to make transgenics for a particular species, the simplicity of the approach allows lineaging to be performed directly on any transgenic line, reducing the time required to perform an experiment. In a similar manner, if one wishes to explore whether lineage specification is perturbed in human IPS cell-based organoid models of disease, then one can use the same tool in all of the organoids derived from different individuals, rather than having to generate a new transgenic for each one. Third, our approach can be used in combination with single cell transcriptomics if the G&T pipeline is applied. This could potentially result in all transcriptome sequenced cells having a corresponding lineage barcode, as one could limit the single cell transcriptomics to cells that have first been confirmed to have their lineage bar code(s) recovered. Fourth, our approach is easily scalable if longer read sequences can be used to add more sites to an array or when it is used in a single cell sequencing pipeline so that one can simply keep adding more arrays to get the number of desired sites for their lineaging experiment. Finally, the allelic sequencing allows the generation of multiple dendrograms for each individual embryo, which promises to increase the amount of information and/or improve the accuracy of lineaging when used in combination with other information such as cell type or cell spatial position.

To conclude, the advent of CRISPR technology has permitted the development of efficient whole-organism lineage-tracing tools. The method described here is the first to use only endogenous CRISPR/Cas9 sequences from the wild-type genome, thus dramatically simplifying the procedure, and perhaps more importantly making whole-organism lineage tracing more feasible in non-model species for which it is hard to generate transgenic animals.

MATERIALS AND METHODS

Scanning the genome

We obtained the repeat masked genomes of zebrafish (danRer11), mouse (mm10) and human (hg38) from the University of California Santa Cruz (UCSC) genome browser. In order to focus on sequences whose genomic origin we could accurately identify and target, we discounted the ChrUn, ChrM and ChrN_random data. Genomes were scanned using a moving window of 450 bp. The first and last 50 bp of each window was reserved for primer identification. The remaining central 350 bp was analyzed for the presence of non-overlapping CRISPR/Cas9 target sites.

Initial filters

At least nine non-overlapping sites consisting of (21xN)GG or CC(21xN) had to be identified in order for the window to be processed further. Base frequency was quantified in the window and a filter included that only allowed the processing of windows where the most frequent base was, at most, 50% more common than the least frequent base.

Assaying off targets

Selected arrays were then further filtered by off target analysis. The set of potential CRISPR/Cas9 target sites (the 20 bp corresponding to the targeting region) from each array was then queried with Bowtie2. Bowtie2 used a prebuilt index generated from the repeat masked zebrafish, mouse and human genomes (danRer11, mm10 and hg38, respectively). The Bowtie2 parameters used were *N*=1 and k=2, otherwise default. After calling bowtie, we analyzed the number of alignment results. If that number was above the number of CRISPR/Cas9 sites queried, then the array failed the filter as this means that there were at least two successful alignments for at least one of the CRISPR/Cas9 sites.

Assaying for likely functional regions

Selected arrays were then assayed to see whether they were likely to inhabit functional regions of the genome. We downloaded the mRNA and upstream5000 files of the zebrafish (danRer10 – these files are not available for danRer11), mouse (mm10) and human (hg38) genomes from the UCSC genome browser. We combined these two files into one and constructed a bowtie2 index for each respective genome. We then called bowtie2 using the entire window of interest as a query sequence. The only Bowtie2 flag used was k=2. If the window did not align to anywhere in the functional region then it passed the filter.

Assaying for regions that likely harbor single nucleotide polymorphisms

In order to select for regions of the genome that likely harbor SNPs, we downloaded from the common SNP track of the UCSC table browser for the mouse (Common SNPS 142) and human (Common SNPS 150) genomes in bed format. As this track does not exist for zebrafish, we downloaded the SNP data from the Zebrafish Information Network (https://zfin.org/downloads/snpData.txt). The CRISPR/Cas9 search algorithm uses this bed file to confirm that the array contain at least one common SNP.

Validation of CRISPR/Cas9 targeting in a mouse or human cell line

Targeting vectors were generated by ligating annealed oligonucleotides corresponding to the sense and antisense of the target region into PX458 following the protocol of Ran et al. (2013). Cell lines were obtained from the cell line repository of the Biomolecular Screening and Protein Technologies Unit (Centre for Genomic Regulation) where they are authenticated and confirmed contaminant free. NIH3T3 cells were transfected with the appropriate plasmid and Lipofectamine 2000, according to the manufacturer's instructions. HEK293 were transfected using 2500 ng PX458 plasmid/well of a six-well plate. Each plasmid was diluted into 250 mM $CaCl_2$ and then an equivolume of 2× HBS [50 mM HEPES, 280 mM NaCl and 1.5 mM Na_2HPO_4 (pH7.0)] was added drop wise and incubated for 30 min at room temperature. This solution was then added dropwise to the cells. For the surveyor nuclease assay, genomic DNA was extracted 65 h after transfection with the Qiagen Blood and Tissue kit. For the Miseq deep-sequencing experiment, cells were first selected for

GFP expression by FACS (using a BD FACSAria II SORP or BD Influx cell sorter) before extraction of genomic DNA. Genomic DNA was then further interrogated by surveyor nuclease assay or Miseq deep sequencing, as described below. For both assays, genomic DNA was extracted from ∼3 million cells.

Transcription of sgRNAs for zebrafish and mouse embryo microinjection

Targeting vectors were generated by annealing and extending the sense and antisense sgRNA targeting oligonucleotides given in Tables S7 and S8 using Phusion polymerase (98°C for 2 min, 50°C for 10 min and 72°C for 10 min). The NEB Hiscribe T7 transcription kit was then used to generate sgRNA (16 h at 37°C) followed by the addition of DNase I for 15 min at 37°C. sgRNA was then purified using Zymogen clean and concentrator kit.

Validation of CRISPR/Cas9 targeting and lineage tracing in zebrafish

sgRNA was microinjected at 100 ng/µl into one-cell stage zebrafish embryo yolk sacs (AB and TL strains) with 8 µM EnGen Cas9 (NEB) and 50 mM KCl, 3 mM $MgCl_2$, 5 mM Tris HCl (pH 8.0) and 0.05% Phenol Red. For validation of CRISPR/Cas9 targeting, sgRNAs were microinjected individually and genomic DNA extracted 30 h later from five individual zebrafish embryos using the Qiagen blood and tissue kit. For lineage tracing, the Z4 sgRNAs were pooled and microinjected, and genomic DNA extracted from five individuals 48 h later using the Qiagen blood and tissue kit.

Validation of CRISPR/Cas9 targeting and lineage tracing in mouse embryos

All experimental protocols were performed in accordance with recommendations for the proper care and use of laboratory animals [local (law 32/2007); European (EU directive 86/609, EU decree 2001-486) regulations and the Standards for Use of Laboratory Animals A5388-01 (NIH)], and were approved by the local ethical committee (CEEA-PRBB).

B6CBAF1 (C57BL/6XCBA) female mice (6-12 weeks) were superovulated by injecting 5 IU PMSG and hCG 47 h later. Females were immediately mated to B6CBAF1 males. Zygotes were collected 20 h post hCG injection. Zygotes were injected into the cytoplasm with the RNP complex consisting of 50 ng/µl Cas9 protein (PNABio CP-02) and a pool of the M7 sgRNAs containing 25 ng/µl of each of the individual sgRNAs. Zygotes were transferred into pre-equilibrated 5 cm dishes containing drops of KSOM+aa and covered by embryo-tested mineral oil. For experiments at later stages of development, 3.5 dpc blastoysts were transferred into the uterine horns of 2.5 dpc pseudopregnant CD1 females and allowed to implant and develop further. Embryos were harvested at E9.0 to E10.0 for further analyses. Genomic DNA was extracted from pools of embryos at the blastocyst stage or individual embryos at the E9.0-E10.0 stage using the Qiagen Blood and Tissue kit.

Surveyor nuclease assay

Amplicons were amplified from genomic DNA using RedTaq polymerase (Sigma). For M7 and H5, we used an amplicon that was larger than our target array for this assay as this results in larger DNA fragments that are easier to detect (see Table S11). Owing to imperfect transfection, we expected amplicons to consist of a mix between wild-type (reference) sequences and mutated sequences. The hybrid mixes of amplicons were thus annealed by heating to 95°C and gradually ramping down to 25°C according to the surveyor nuclease protocol. Surveyor nuclease, enhancer and $MgCl_2$ were then added to the hybridized PCR products in the volumetric ratio of 1:1:0.6:6. The surveyor nuclease reaction was then carried out at 42°C for 1 h. The resulting reaction was stopped with 1/10 volume of STOP solution and the product cleaned using the Nucleospin PCR cleanup kit. DNA was eluted in 15 µl of water and 1 µl was run on a high-sensitivity DNA chip (Agilent). Controls were run as suggested by the manufacturer.

Miseq deep sequencing

We amplified the Z4, M7 and H5 amplicons with two rounds of PCR (20-35 and 15-20 cycles) using the Q5 high-fidelity polymerase (NEBNext 2x PCR master mix). The first round used the genomic DNA extracted from the 3T3, HEK293 cells mouse blastocysts/embryos and zebrafish embryos using the standard Z4, M7 and H5 primers identified in our bioinformatic analysis (see Table S11). The second round was used to directly add the Illumina adapters to these amplicons (see Table S12). Furthermore, we added random base pairs (5xN for version 1 and 16xN for version 2) directly 3′ to the Illumina universal primer to aid Illumina cluster resolution and to serve as a unique molecular identifier (UMI). Adapted amplicons were pooled and sequenced on the Illumina Miseq using the version 2 or 3 sequencing kit at 2×250 cycles for version 1 of the pipeline and 2×300 cycles for version 2 of the pipeline.

Bioinformatic processing of reads

We first trimmed poor quality reads using trimmomatic (Bolger et al., 2014) with the following parameters:

"TRAILING:15 SLIDINGWINDOW:4:20 MINLEN:45" for version 1 and

"TRAILING:15 SLIDINGWINDOW:4:20 MINLEN:66" for version 2.

Paired end reads were then amalgamated using PEAR (Zhang et al., 2013). For version 2, we then used an in-house program to remove PCR duplicates and reads that could have potentially been derived from PCR flip-over [described in greater detail in the supplementary Materials and Methods (section S4) and Fig. S2]. Briefly, any identical reads containing the same UMI were collapsed to a single read as these are PCR duplicates. Different reads containing the same UMI were all removed as these are likely derived from a PCR flip-over event. The 5′ random bases were then removed with cutadapt (Martin, 2011).

Quantifying editing efficiency

Processed reads were mapped to the mouse (mm10), zebrafish (danRer11) or human (hg38) genomes using bowtie2 with default parameters (end-to-end and sensitive mode, rdg=5 3 and rfg=5 3). We wrote an in-house program that processes the CIGAR string of the mapped reads. Specifically, it uses the CIGAR string to calculate at which position of the mapped read an edit occurs. The midpoint of an edit is used to record the position of any particular edit. For example, a CIGAR string of 195M5D200M would record an edit at position 198. These data are normalized to the number of reads mapped to that particular genomic site to measure editing efficiency, as shown in Fig. 3. Furthermore, the size and the types of the different edits are recorded along with any deletions that span from any one spacer region to the next (i.e. start point in one spacer and end point in the adjacent spacer). These data are used to generate the panels shown in Fig. S3.

Processing Illumina sequence data for zebrafish dendrogram construction

Processed reads were mapped to the zebrafish (danRer11) genome using the bowtie2 and Bwa mem aligners. Reads were also mapped to the zebrafish Z4 amplicon using the Needleall aligner [full details of how the resulting fasta file was processed are given in the supplementary Material and Methods (section S5)]. Only reads that had the same reported CIGAR string with all three aligners were kept to generate a high confidence dataset for dendrogram construction. Three sets of parameters were used for these data:

1. Bowtie2 (rdg=5 3 and rfg=5 3, other parameters as default), Bwa mem (O=5, E=3, w=400, other parameters as default), Needleall (gapopen=5, gapextend=3, other parameters as default).
2. Bowtie2 (rdg=6 1 and rfg=6 1, other parameters as default), Bwa mem (O=6, E=1, w=400, other parameters as default), Needleall (gapopen=6, gapextend=1, other parameters as default).
3. Bowtie2 (rdg=10 1 and rfg=10 1, other parameters as default), Bwa mem (O=10, E=0.5, w=400, other parameters as default), Needleall (gapopen=10, gapextend=0.5, other parameters as default).

Processing Illumina sequence data for mouse dendrogram construction

Owing to the lower amount of sequence diversity generated in the M7 array, we used only a single aligner for the mouse data. Processed reads were mapped to the mouse genome (mm10) using bwa mem with parameters O=6, E=1 and w=400 (all other parameters as default). We then applied an in-house c-program that used the MD flag of the SAM format file to determine the allele of the M7 read. The program used the MD string to interrogate positions 189 and 191 of the M7 array to determine whether they are the TTCC or CTCT haplotype allele type and to export the specific mapping to separate allele SAM format files [see supplementary Materials and Methods (section S7) and Table S6 for more details).

Construction of dendrogram and quantification of dendrogram depth

We first defined a feature array for all reads by scanning through the CIGAR string of each mapped read and defining which particular combination of indels occurs within the 10 spacer regions [see Fig. S8 and supplementary Materials and Methods (section S6 for more details]. The spacer regions of the Z4 and M7 arrays used to generate the feature array are listed in Tables S4 and S5. The exact position of the feature within the window was irrelevant, which helps with offset differences between reads due to other mutations (which can generate staggered but otherwise identical edits). We only considered indels >1 bp to reduce the possibility that sequencing errors could add noise to the dendrogram.

We used PAUP* (version 4.0a build 166) (Wang et al., 2013) to construct lineage trees with the feature array data in a similar manner to that described by Salvador-Martínez et al. (2019) and McKenna et al. (2016). However, we implemented PAUP* with the full alphanumerical set of 36 character states per site (using the 36 most frequent mutations at each site) and generated a different triangular distance matrix for each target site in the array. The distance matrices for each target site were generated using the following criteria:

$$d(i,j) = 0, \quad \text{if } i = j$$

$$d(i,j) = \log\frac{1}{(freq_j)}, \text{ if } i \text{ is unmutated}$$

$$d(i,j) = \log\frac{1}{(freq_i)}, \text{ if } j \text{ is unmutated and}$$

$$d(i,j) = \log\frac{1}{(freq_j)} + \log\frac{1}{(freq_i)} \quad \text{otherwise,}$$

where *freq* is the frequency of a given mutation *i* or *j* in a given dataset. Owing to the large number of alleles in our zebrafish data, we used PAUP* in neighbor joining mode. For the purposes of consistency, we also used PAUP* in neighbor joining mode for mouse lineaging. We used IcyTree (Vaughan, 2017) to plot the dendrograms with zero-length edges collapsed.

Acknowledgements

We thank Jan Philipp Junker, Irepan Salvador-Martínez, Cedric Notredame, Emilio Palumbo and Ben Lehner for helpful comments and suggestions. Furthermore, we thank Antoni Matyjaszkiewicz for proof reading. We thank the UPF Genomics Core Facility for performing the Illumina Miseq deep sequencing, Sanger sequencing and bioanalyzer analysis. We also thank the PRBB Animal facility for Zebrafish and Mouse husbandry. We also thank the UPF Flow Cytometry core facility for performing all cell sorting.

Competing interests

The authors declare no competing or financial interests.

Author contributions

Conceptualization: J.C., J.S.; Methodology: J.C., M.V.C., L.B.M.; Software: J.C.; Validation: J.C.; Formal analysis: J.C.; Investigation: J.C., M.V.C., L.B.M.; Resources: J.C.; Data curation: J.C.; Writing - original draft: J.C., J.S.; Writing - review & editing: J.C., J.S.; Visualization: J.C.; Supervision: J.S.; Project administration: J.C., J.S.; Funding acquisition: J.S.

Funding

This work was funded by the core program funding of the European Molecular Biology Laboratory with support from the European Research Council through SIMBIONT (670555). Deposited in PMC for immediate release.

Data availability

Raw fastq datasets used in this study can be found on the European Nucleotide Archive under accession number PRJEB37916.

Supplementary information

Supplementary information available online at
http://dev.biologists.org/lookup/doi/10.1242/dev.184481.supplemental

Peer review history

The peer review history is available online at
https://dev.biologists.org/lookup/doi/10.1242/dev.184481.reviewer-comments.pdf

References

Alemany, A., Florescu, M., Baron, C. S., Peterson-Maduro, J. and Van Oudenaarden, A. (2018). Whole-organism clone tracing using single-cell sequencing. *Nature* **556**, 108-112. doi:10.1038/nature25969

Bolger, A. M., Lohse, M. and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina Sequence Data. *Bioinformatics* **30**, 2114-2120. doi:10.1093/bioinformatics/btu170

Briggs, J. A., Weinreb, C., Wagner, D. E., Megason, S., Peshkin, L., Kirschner, M. W. and Klein, A. M. (2018). The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360**, eaar5780. doi:10.1126/science.aar5780

Burstein, D., Harrington, L. B., Strutt, S. C., Probst, A. J., Anantharaman, K., Thomas, B. C., Doudna, J. A. and Banfield, J. F. (2017). New CRISPR-Cas systems from uncultivated microbes. *Nature* **542**, 237-241. doi:10.1038/nature21059

Chari, R., Yeo, N. C., Chavez, A. and Church, G. M. (2017). sgRNA Scorer 2.0: a species-independent model to predict CRISPR/Cas9 activity. *ACS Synth. Biol.* **6**, 902-904. doi:10.1021/acssynbio.6b00343

Clarke, J. D. W. and Tickle, C. (1999). Fate maps old and new. *Nat. Cell. Biol* **1**, E103-E109. doi:10.1038/12105

Doench, J. G., Hartenian, E., Graham, D. B., Tothova, Z., Hegde, M., Smith, I., Sullender, M., Ebert, B. L., Xavier, R. J. and Root, D. E. (2014). Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat. Biotechnol.* **32**, 1262-1267. doi:10.1038/nbt.3026

Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F., Smith, I., Tothova, Z., Wilen, C., Orchard, R. et al. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184-191. doi:10.1038/nbt.3437

Doudna, J. A. and Charpentier, E. (2014). Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science* **346**, 1258096. doi:10.1126/science.1258096

Farrell, J. A., Wang, Y., Riesenfeld, S. J., Shekhar, K., Regev, A. and Schier, A. F. (2018). Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360**, eaar3131. doi:10.1126/science.aar3131

Frieda, K. L., Linton, J. M., Hormoz, S., Choi, J., Chow, K.-H. K., Singer, Z. S., Budde, M. W., Elowitz, M. B. and Cai, L. (2017). Synthetic recording and in situ readout of lineage information in single cells. *Nature* **541**, 107-111. doi:10.1038/nature20777

Frumkin, D., Wasserstrom, A., Kaplan, S., Feige, U. and Shapiro, E. (2005). Genomic variability within an organism exposes its cell lineage tree. *PLoS Comput. Biol.* **1**, e50. doi:10.1371/journal.pcbi.0010050

Frumkin, D., Wasserstrom, A., Itzkovitz, S., Stern, T., Harmelin, A., Eilam, R., Rechavi, G. and Shapiro, E. (2008). Cell lineage analysis of a mouse tumor. *Cancer Res.* **68**, 5924-5931. doi:10.1158/0008-5472.CAN-07-6216

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A. and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821. doi:10.1126/science.1225829

Junker, J. P., Spanjaard, B., Peterson-Maduro, J., Alemany, A., Hu, B., Florescu, M. and Van Oudenaarden, A. (2017). Massively parallel clonal analysis using CRISPR/Cas9 induced genetic scars. *bioRxiv*. doi:10.1101/056499

Kalhor, R., Mali, P. and Church, G. M. (2017). Rapidly evolving homing CRISPR barcodes. *Nat. Methods.* **14**, 195-200. doi:10.1038/nmeth.4108

Kalhor, R., Kalhor, K., Mejia, L., Leeper, K., Graveline, A., Mali, P. and Church, G. M. (2018). Developmental barcoding of whole mouse via homing CRISPR. *Science* **361**, eaat9804. doi:10.1126/science.aat9804

Kleinstiver, B. P., Prew, M. S., Tsai, S. Q., Topkar, V. V., Nguyen, N. T., Zheng, Z., Gonzales, A. P. W., Li, Z., Peterson, R. T., Yeh, J.-R. J. et al. (2015). Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* **523**, 481-485. doi:10.1038/nature14592

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357-359. doi:10.1038/nmeth.1923

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. [arXiv:1303.3997v2][12] [q-bio.GN].

Livet, J., Weissman, T. A., Kang, H., Draft, R. W., Lu, J., Bennis, R. A., Sanes, J. R. and Lichtman, J. W. (2007). Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature* **450**, 56-62. doi:10.1038/nature06293

Macaulay, I. C., Haerty, W., Kumar, P., Li, Y. I., Hu, T. X., Teng, M. J., Goolam, M., Saurat, N., Coupland, P., Shirley, L. M. et al. (2015). G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12**, 519-522. doi:10.1038/nmeth.3370

Mao, X., Fujiwara, Y., Chapdelaine, A., Yang, H. and Orkin, S. H. (2001). Activation of EGFP expression by Cre-mediated excision in a new ROSA26 reporter mouse strain. *Blood* **97**, 324-326. doi:10.1182/blood.V97.1.324

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* **17**. 1, 10-12. doi:10.14806/ej.17.1.200

Mckenna, A., Findlay, G. M., Gagnon, J. A., Horwitz, M. S., Schier, A. F. and Shendure, J. (2016). Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907. doi:10.1126/science.aaf7907

Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. and Almendros, C. (2009). Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**, 733-740. doi:10.1099/mic.0.023960-0

Prolla, T. A., Baker, S. M., Harris, A. C., Tsao, J.-L., Yao, X., Bronner, C. E., Zheng, B., Gordon, M., Reneker, J., Arnheim, N. et al. (1998). Tumour susceptibility and spontaneous mutation in mice deficient in Mlh1, Pms1 and Pms2 DNA mismatch repair. *Nat. Genet.* **18**, 276-279. doi:10.1038/ng0398-276

Raj, B., Wagner, D. E., Mckenna, A., Pandey, S., Klein, A. M., Shendure, J., Gagnon, J. A. and Schier, A. F. (2018). Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36**, 442-450. doi:10.1038/nbt.4103

Ran, F. A., Hsu, P. D., Wright, J., Agarwala, V., Scott, D. A. and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281-2308. doi:10.1038/nprot.2013.143

Rice, P., Longden, I. and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276-277. doi:10.1016/S0168-9525(00)02024-2

Salipante, S. J. and Horwitz, M. S. (2006). Phylogenetic fate mapping. *Proc. Natl. Acad. Sci. USA* **103**, 5448-5453. doi:10.1073/pnas.0601265103

Salvador-Martínez, I., Grillo, M., Averof, M. and Telford, M. J. (2019). Is it possible to reconstruct an accurate cell lineage using CRISPR recorders? *eLife* **8**, e40292. doi:10.7554/eLife.40292

Schmidt, S. T., Zimmerman, S. M., Wang, J., Kim, S. K. and Quake, S. R. (2017). Quantitative analysis of synthetic cell lineage tracing using nuclease barcoding. *ACS Synth. Biol.* **6**, 936-942. doi:10.1021/acssynbio.6b00309

Shah, S. A., Erdmann, S., Mojica, F. J. M. and Garrett, R. A. (2013). Protospacer recognition motifs: mixed identities and functional diversity. *RNA Biol.* **10**, 891-899. doi:10.4161/rna.23764

Shibata, D., Navidi, W., Salovaara, R., Li, Z.-H. and Aaltonen, L. A. (1996). Somatic microsatellite mutations as molecular tumor clocks. *Nat. Med.* **2**, 676-681. doi:10.1038/nm0696-676

Soriano, P. (1999). Generalized lacZ expression with the ROSA26 Cre reporter strain. *Nat. Genet.* **21**, 70-71. doi:10.1038/5007

Spanjaard, B., Hu, B., Mitic, N., Olivares-Chauvet, P., Janjuha, S., Ninov, N. and Junker, J. P. (2018). Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. *Nat. Biotechnol.* **36**, 469-473. doi:10.1038/nbt.4124

Stern, C. D. and Fraser, S. E. (2001). Tracing the lineage of tracing cell lineages. *Nat. Cell Biol.* **3**, E216-E218. doi:10.1038/ncb0901-e216

Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. and Doudna, J. A. (2014). DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**, 62-67. doi:10.1038/nature13011

Stirparo, G. G., Boroviak, T., Guo, G., Nichols, J., Smith, A. and Bertone, P. (2018). Integrated analysis of single-cell embryo data yields a unified transcriptome signature for the human pre-implantation epiblast. *Development* **145**, dev158501. doi:10.1242/dev.158501

Sulston, J. E., Schierenberg, E., White, J. G. and Thomson, J. N. (1983). The embryonic cell lineage of the nematode Caenorhabditis elegans. *Dev. Biol.* **100**, 64-119. doi:10.1016/0012-1606(83)90201-4

Swofford, D. L. (2017). *PAUP* Phylogenetic Analysis Using Parsimony (*and Other Methods)*, version 4.0a. Sunderland, Massachusetts: Sinauer Associates.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M. and Rozen, S. G. (2012). Primer3—new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115. doi:10.1093/nar/gks596

Vaughan, T. G. (2017). IcyTree: rapid browser-based visualization for phlogenetic trees and networks. *Bioinformatics* **33**, 2392-2394. doi:10.1101/110213

Vogt, W. (1929). Gestaltungsanalyse am Amphibienkeim mit Örtlicher Vitalfärbung. II. Teil Gastrulation und Mesodermbildung bei Urodelen und Anuren. *Wilhelm Roux Arch. Entwicklungsmech. Org.* **120**, 384-706. doi:10.1007/BF02109667

Wagner, D. E., Weinreb, C., Collins, Z. M., Briggs, J. A., Megason, S. G. and Klein, A. M. (2018). Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981-987. doi:10.1126/science.aar4362

Wang, H., Yang, H., Shivalila, C. S., Dawlaty, M. M., Cheng, A. W., Zhang, F. and Jaenisch, R. (2013). One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153**, 910-918. doi:10.1016/j.cell.2013.04.025

Wang, T., Wei, J. J., Sabatini, D. M. and Lander, E. S. (2014). Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80-84. doi:10.1126/science.1246981

Weisblat, D. A., Sawyer, R. T. and Stent, G. S. (1978). Cell lineage analysis by intracellular injection of a tracer enzyme. *Science* **202**, 1295-1298. doi:10.1126/science.725606

Xu, H., Xiao, T., Chen, C.-H., Li, W., Meyer, C. A., Wu, Q., Wu, D., Cong, L., Zhang, F., Liu, J. S. et al. (2015). Sequence determinants of improved CRISPR sgRNA design. *Genome Res.* **25**, 1147-1157. doi:10.1101/gr.191452.115

Zetsche, B., Gootenberg, J. S., Abudayyeh, O. O., Slaymaker, I. M., Makarova, K. S., Essletzbichler, P., Volz, S. E., Joung, J., Van der Oost, J., Regev, A. et al. (2015). Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* **163**, 759-771. doi:10.1016/j.cell.2015.09.038

Zhang, J., Kobert, K., Flouri, T. and Stamatakis, A. (2013). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614-620. doi:10.1093/bioinformatics/btt593